SMDM PROJECT

REPORT

Contents

Problem 1 4

1.1. Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least? 4

1.2. There are 6 different varieties of items that are considered. Describe and comment/explain allthe varieties across Region and Channel? Provide a detailed justification for your answer.....6

1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?.....12

1.4. Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.....12

1.5. On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.....14

Problem 2.....15

2.1. For this data, construct the following contingency tables (Keep Gender as row variable).....13

2.1.1. Gender and Major.....15

2.1.2. Gender and Grad Intention.....15

2.1.3. Gender and Employment.....15

2.1.4. Gender and Computer.....16

2.2. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following questions:.....16

2.2.1. What is the probability that a randomly selected CMSU student will be male? 16

2.2.2.What is the probability that a randomly selected CMSU student will be female?. 16

2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 17

2.3.1. Find the conditional probability of different majors among the male students in CMSU.....17

2.3.2 Find the conditional probability of different majors among the female students of CMSU.....17

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.....

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.....

2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: ……..18

2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment?.....18

2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management……..20

2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?.....21

.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

Problem 3……..24

3.1 Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test forequality of means is performed?.....24

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?.....27.

Problem 1

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data (WholesaleCustomer.csv) consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

1.1. Use methods of descriptive statistics to summarize data.

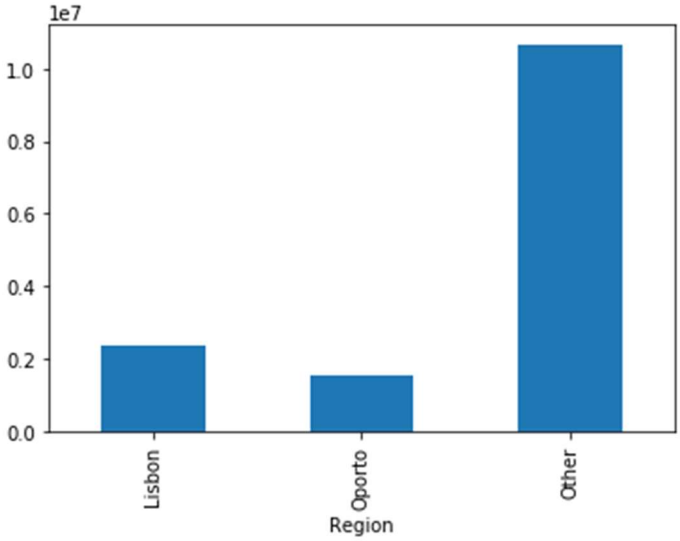
Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	
count	440.00	440.00	440.00	440.00	440.00	440.00	440.00
mean	220.50	12000.30	5796.27	7951.28	3071.93	2881.49	1524.87
std	127.16	12647.33	7380.38	9503.16	4854.67	4767.85	2820.11
min	1.00	3.00	55.00	3.00	25.00	3.00	3.00
25%	110.75	3127.75	1533.00	2153.00	742.25	256.75	408.25
50%	220.50	8504.00	3627.00	4755.50	1526.00	816.50	965.50
75%	330.25	16933.75	7190.25	10655.75	3554.25	3922.00	1820.25
max	440.00	112151.00	73498.00	92780.00	60869.00	40827.00	47943.0

	Channel	Region
count	440	440
unique	2	3
top	Hotel	Other
freq	298	316

1.1. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

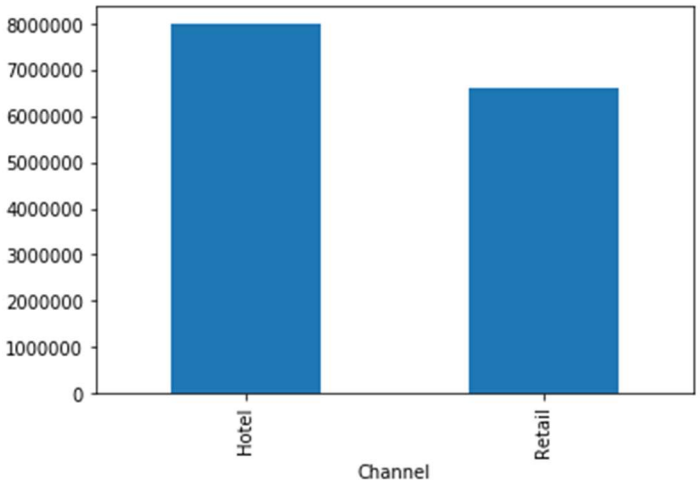
Region	
Lisbon	2386813
Oporto	1555088
Other	10677599

Name: spending, dtype: int64



Highest spend in the Region = Other (10677599) and lowest spend in the region = Oporto (1555088)¶

Channel
Hotel 7999569
Retail 6619931
Name: spending, dtype: int64



Highest spend in the Channel = Hotel(7999569) and lowest spend in the Channel = Retail(6619931).

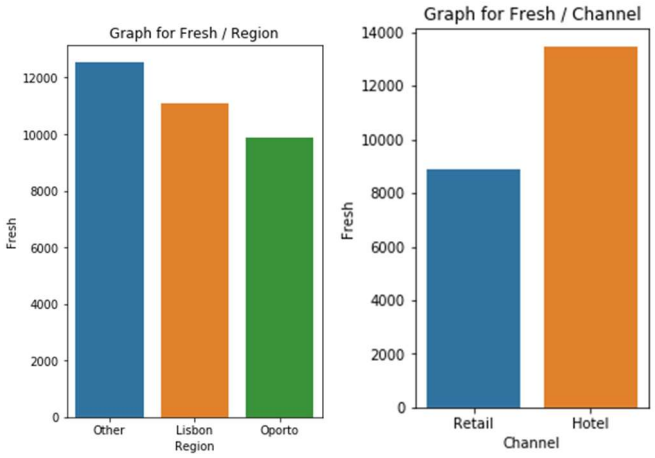
1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

For Fresh

Region
Lisbon 854833
Oporto 464721
Other 3960577
Name: Fresh, dtype: int64
Text(0.5, 1.0, 'Graph for Fresh / Region')

Channel
Hotel 4015717
Retail 1264414
Name: Fresh, dtype: int64

Text(0.5, 1.0, 'Graph for Fresh / Channel')



Comments

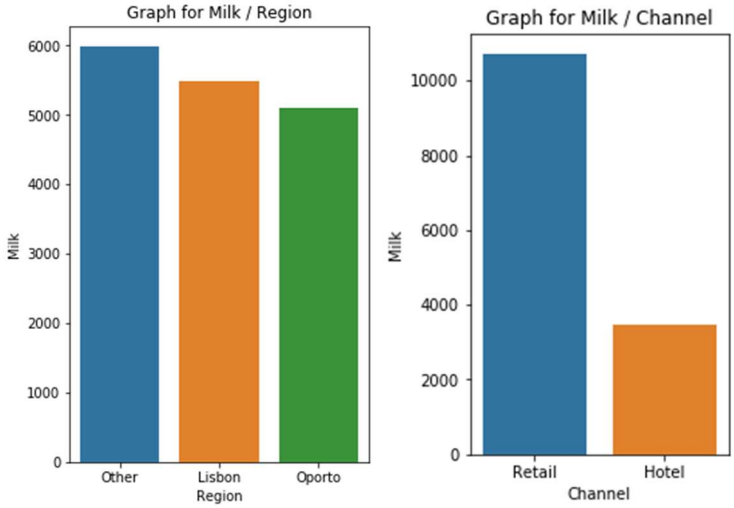
The Fresh is highest spended in the region of "Other"
The Fresh is lowest spended in the region of "Oporto "
The Fresh is highest spended in the Channel of "Hotel"
The Fresh is lowest spended in the Channel of "Retail"

For Milk

Region
Lisbon 422454
Oporto 239144
Other 1888759
Name: Milk, dtype: int64
Text(0.5, 1.0, 'Graph for Milk / Region')

Channel
Hotel 1028614
Retail 1521743
Name: Milk, dtype: int64

Text (0.5, 1.0, 'Graph for Milk / Channel')



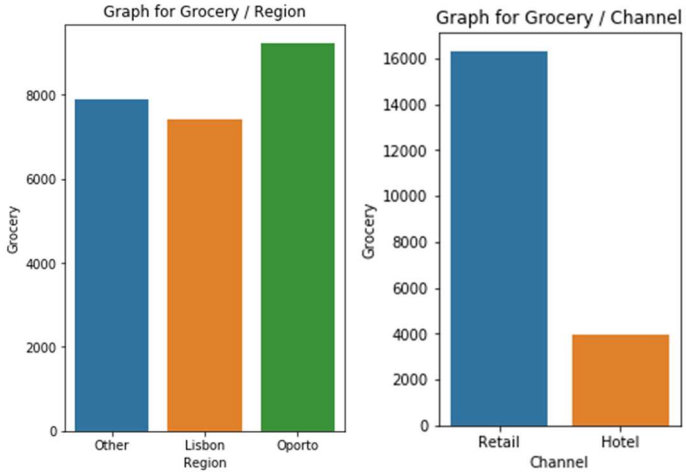
Comments

The Milk is highest spent in the region of "Other"
The Milk is lowest spent in the region of "Oporto "
The Milk is highest spent in the Channel of "Retail"
The Milk is lowest spent in the Channel of "Hotel "

For Grocery

Region
Lisbon 570037
Oporto 433274
Other 2495251
Name: Grocery, dtype: int64
Text(0.5, 1.0, 'Graph for Grocery / Region')

Channel
Hotel 1180717
Retail 2317845
Name: Grocery, dtype: int64



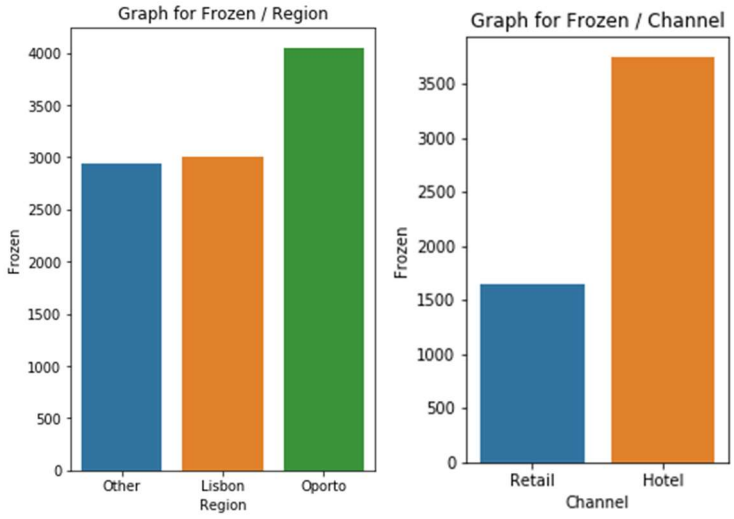
The Grocery is highest spended in the region of "Other"
The Grocery is lowest spended in the region of "Oporto "
The Grocery is highest spended in the Channel of "Retail"
The Grocery is lowest spended in the Channel of "Hotel "

Frozen

Region
Lisbon 231026
Oporto 190132
Other 930492
Name: Frozen, dtype: int64
Text(0.5, 1.0, 'Graph for Frozen / Region')

Channel
Hotel 1116979
Retail 234671
Name: Frozen, dtype: int64

Text(0.5, 1.0, 'Graph for Frozen / Channel')



Comments

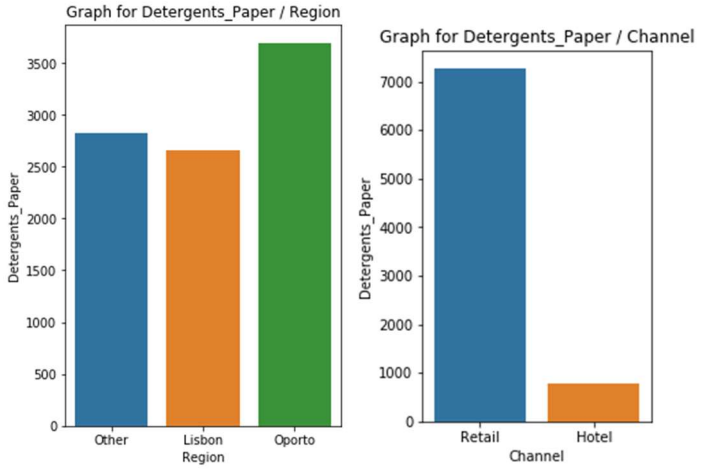
The Frozen is highest spent in the region of "Oporto"
The Frozen is lowest spent in the region of " Others"
The Frozen is highest spent in the Channel of "Hotel"
The Frozen is lowest spent in the Channel of "Retail

Detergents_Paper

Region
Lisbon 204136
Oporto 173311
Other 890410
Name: Detergents_Paper, dtype: int64
Text(0.5, 1.0, 'Graph for Detergents_Paper / Region')

Channel
Hotel 235587
Retail 1032270
Name: Detergents_Paper, dtype: int64

Text(0.5, 1.0, 'Graph for Detergents_Paper / Channel')



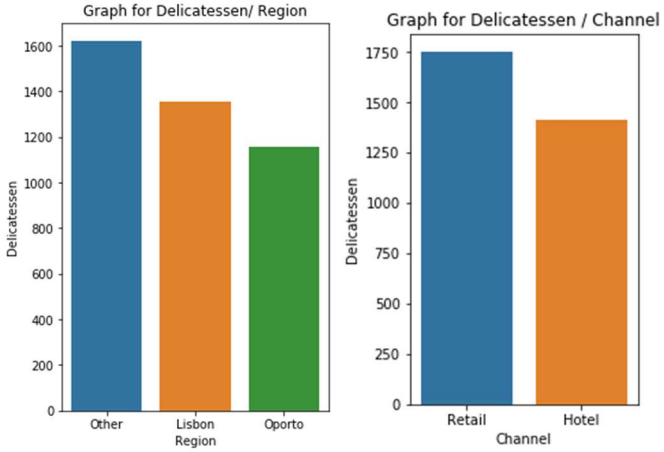
The Detergents_Paper is highest spent in the region of "Oporto"
The Detergents_Paper is lowest spent in the region of " Lisbon"
The Detergents_Paper is highest spent in the Channel of "Retail"
The Detergents_Paper is lowest spent in the Channel of "Hotel"

Delicatessen

Region
Lisbon 104327
Oporto 54506
Other 512110
Name: Delicatessen, dtype: int64
Text(0.5, 1.0, 'Graph for Delicatessen/ Region')

Channel
Hotel 421955
Retail 248988
Name: Delicatessen, dtype: int64

Text(0.5, 1.0, 'Graph for Delicatessen / Channel')



The Delicatessen is highest spent in the region of "Others"
The Delicatessen is lowest spent in the region of " Oporto"
The Delicatessen is highest spent in the Channel of "Retail"
The Delicatessen is lowest spent in the Channel of "Hotel"

1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour?

Which items shows the least inconsistent behaviour?

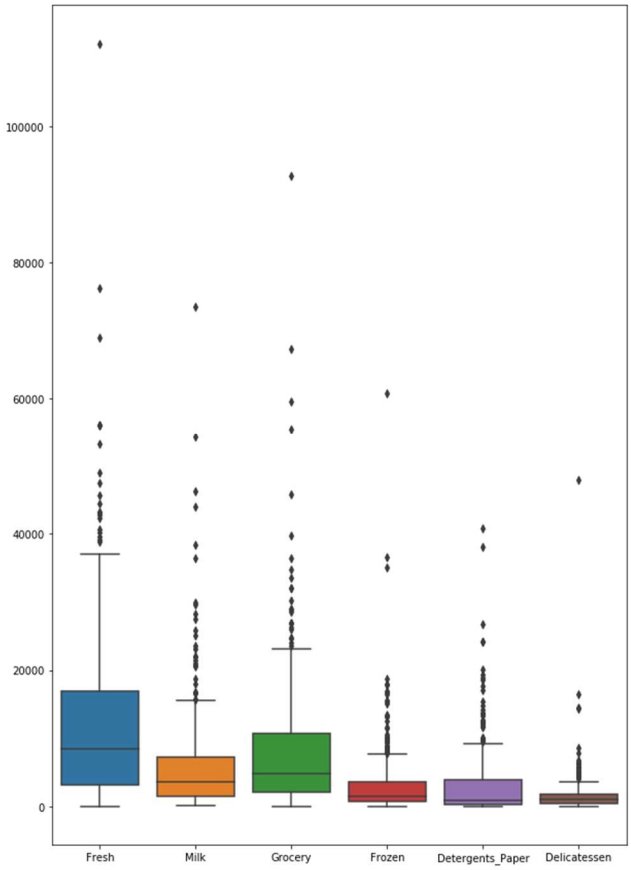
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.00	440.00	440.00	440.00	440.00	440.00
mean	12000.30	5796.27	7951.28	3071.93	2881.49	1524.87
std	12647.33	7380.38	9503.16	4854.67	4767.85	2820.11
min	3.00	55.00	3.00	25.00	3.00	3.00
25%	3127.75	1533.00	2153.00	742.25	256.75	408.25
50%	8504.00	3627.00	4755.50	1526.00	816.50	965.50
75%	16933.75	7190.25	10655.75	3554.25	3922.00	1820.25
max	112151.00	73498.00	92780.00	60869.00	40827.00	47943.00
Total	5280131.00	2550357.00	3498562.00	1351650.00	1267857.00	670943.00
CV	1.05	1.27	1.20	1.58	1.65	1.85

From the above table, we can say that, 'Frozen' items show the highest inconsistent behaviour in as the coefficient of variation is the high for it. On the other hand, 'Fresh' show the most consistent behaviour as the CV is the lowest.

1.4. Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	12669	9656	7561	214	2674	1338
1	7057	9810	9568	1762	3293	1776
2	6353	8808	7684	2405	3516	7844
3	13265	1196	4221	6404	507	1788
4	22615	5410	7198	3915	1777	5185
...	...	...	...	...	...	...

435	29703	12051	16027	13135	182	2204
436	39228	1431	764	4510	93	2346
437	14531	15488	30243	437	14841	1867
438	10290	1981	2232	1038	168	2125
439	2787	1698	2510	65	477	52



Box plot is the best pictorial presentation to show the outliers. All products have outliers.

1.5. On the basis of your analysis, what are your recommendations for the business?How can your analysis help the business to solve its problem? Answer from the business perspective.

Highest spend in the Region = Other (10677599) and
lowest spend in the region = Oporto (1555088)¶
Highest spend in the Channel = Hotel(7999569) and
lowest spend in the Channel = Retail(6619931).

'Frozen' products iteams show the highest inconsistent behavior
'Fresh' products show the most consistent behavior

In Channel "Hotel" Average Highest Spending in Fresh items and Lowest Spending in Detergents_Paper.¶
In Channel "Retail" Average Highest Spending in Grocery items and Lowest Spending in Frozen items.

In Region "Lisbon" Average Highest Spending in Fresh and Lowest in Delicatessen items.
In Region "Oporto" Average Highest Spending in Fresh and Lowest in Delicatessen items.
In Region "Other" Average Highest Spending in Fresh and Lowest in Delicatessen items.

Problem 2

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14questions and receives responses from 62 undergraduates (stored in the Survey.csv file).

2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major

Major	Accounting	CIS	Economics/ Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

2.1.2. Gender and Grad Intention

Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

2.1.3. Gender and Employment

Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	3	24	6
Male	7	19	3

2.1.4. Gender and Computer

Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

2.2. Assume that the sample is a representative of the population of CMSU. Based onthe data, answer the following questions:

2.2.1. What is the probability that a randomly selected CMSU student will be male?

Female 33
Male 29
Name: Gender, dtype: int64

2.2.2. Find the conditional probability of different majors among the female students of CMSU.

Probability that a randomly selected CMSU student will be female: 0.46774193548387094

2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.3.1. Find the conditional probability of different majors among the male students in CMSU.¶¶

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

Conditional probability of Accounting: 0.13793103448275862
Conditional probability of CIS: 0.034482758620689655
Conditional probability of Economics/Finance: 0.13793103448275862
Conditional probability of International Business: 0.06896551724137931
Conditional probability of Management: 0.20689655172413793
Conditional probability of Other: 0.13793103448275862
Conditional probability of Retailing/Marketing: 0.1724137931034483
Conditional probability of Undecided: 0.10344827586206896

2.3.2 Find the conditional probability of different majors among the female students of CMSU.

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

Conditional probability of Accounting: 0.09090909090909091
Conditional probability of CIS: 0.09090909090909091
Conditional probability of Economics/Finance: 0.21212121212121213
Conditional probability of International Business: 0.12121212121212122
Conditional probability of Management: 0.12121212121212122
Conditional probability of Other: 0.09090909090909091
Conditional probability of Retailing/Marketing: 0.2727272727272727
Conditional probability of Undecided: 0.0

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.

Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

P(Graduate ∩ Male) = P (Graduate| Male) x P (male) = 0.27419354838709675

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.

Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

P(Graduate ∩ female) = P (Graduate| female) x P (female) = 0.064516129032258 06

2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment?

P(A U B) = P(A)+P(B)

Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	3	24	6
Male	7	19	3

The probability of being male is 0.46774193548387094
The probability of being Full-time is 0.2413793103448276

probability of being male or being full-time employee= 0.70912124
58286985

2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

P(A U B) = P(A)+P(B)

Major	Accounting	CIS	Economics/ Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

The probability of being female in International Business is 0.125
The probability of being female in Management is 0.125

The probalility female student majoring in international business or management= 0.25

2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

The graduate intention and being female are independent events?

For 2 events to be independent, following condition is to be satified

P(A ∩ B) = P(A) * P(B)

So, P (Graduate intention ∩ Female) = P(Graduate intention) * P(Female)

P(Female) = 0.8
P(Graduate intention) = 0.7

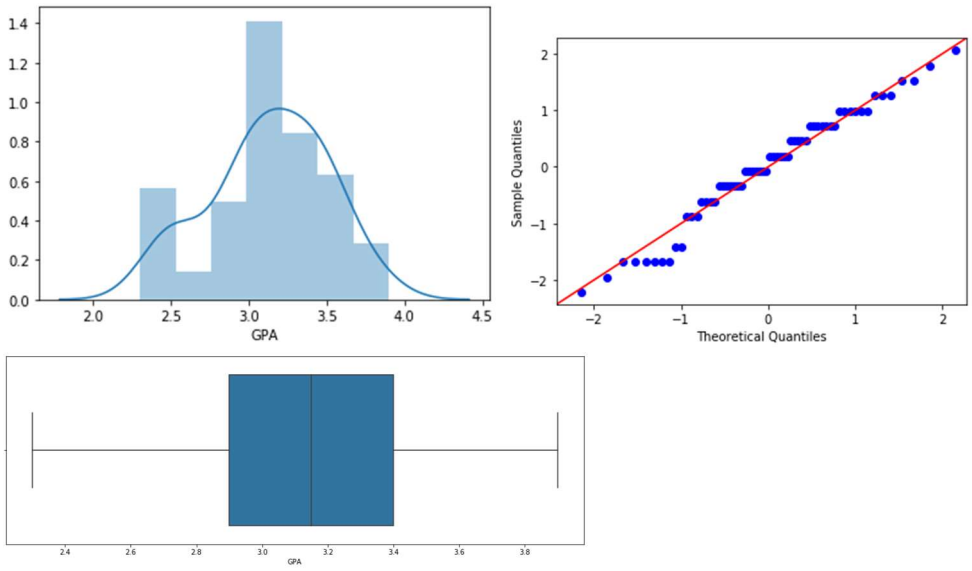
P(Graduate intention) * P(Female) = 0.5599999999999999
P(Graduate intention ∩ Female) = 0.275

This is not independent events as probability multiplication of both events is not equal to com bined event, so being a winner and being female candidate are not independent events

2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

Normality distribuation test GPA

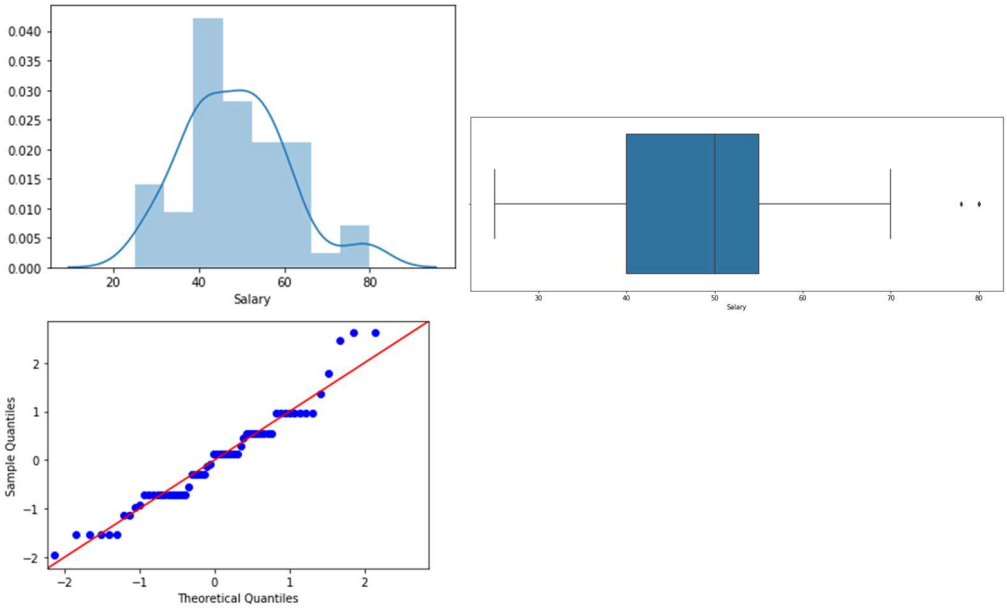
Chi2=1.95, pvalue=0.377



GPA is normally distributed

Normality distribuation test Salary

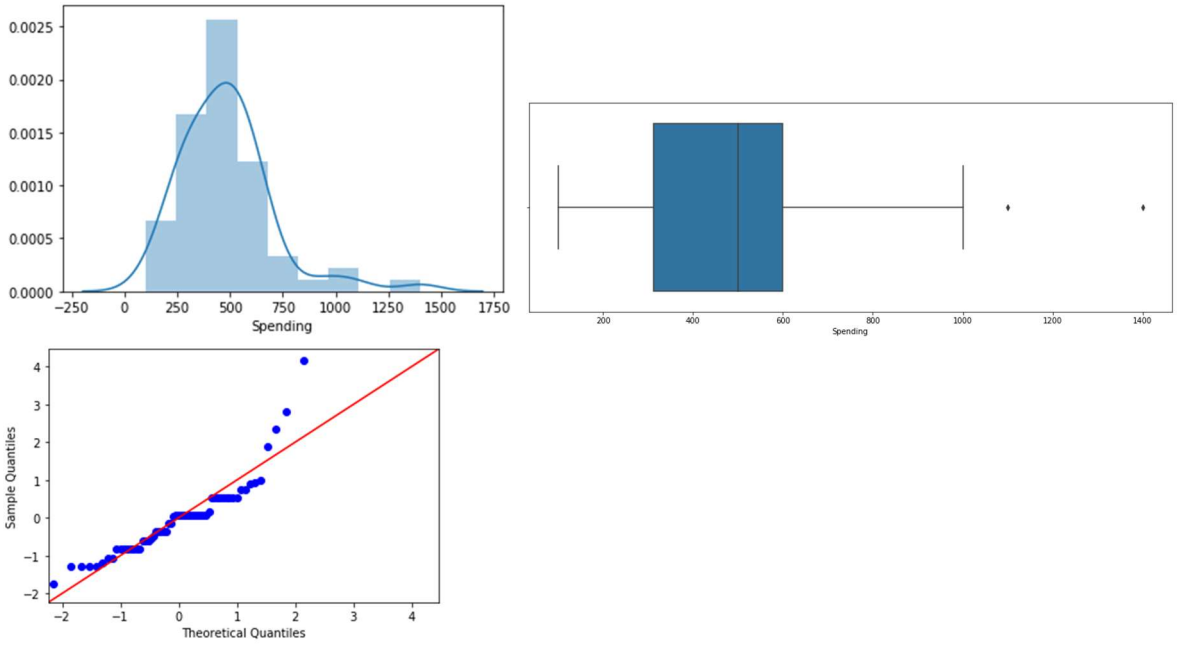
Chi2=3.85, pvalue=0.146



Salary is normally distributed

Normality distribuation test Spending

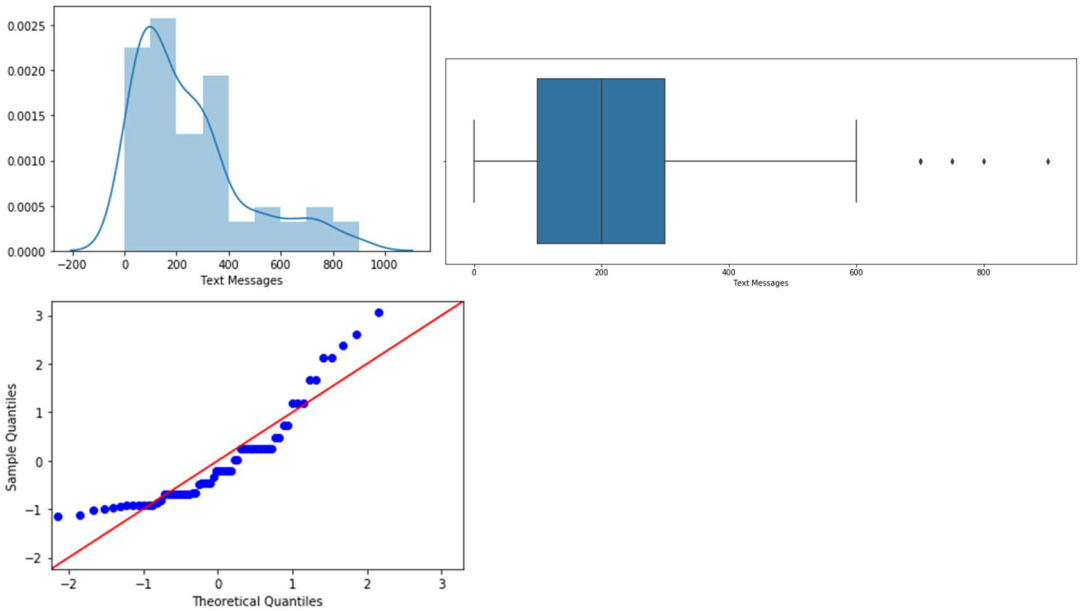
Chi2=30.50, pvalue=0.000



Spending is not normally distributed

Normality distribuation test Text Messages

Chi2=16.35, pvalue=0.000



Text Messages is not normally distributed

Problem 3

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amountof moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the
packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. To monitorthe amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company claims that the mean moisture content cannot be greater than 0.35 pound per 100 square feet.
The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shinglesand 31 for B shingles.

For the A shingles, the null and alternative hypothesis to test whether the population mean moisturecontent is less than 0.35 pound per 100 square feet is given:

$$H_0 \leq 0.35$$

$$H_1 > 0.35$$

For the B shingles, the null and alternative hypothesis to test whether the population mean moisturecontent is less than 0.35 pound per 100 square feet is given:

$$H_0 \leq 0.35$$

$$H_1 > 0.35$$

3.1 Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need tocheck before the test for equality of means is performed?

1 sample T test for shingles 'A'

0	0.44
1	0.61
2	0.47
3	0.30
4	0.15
5	0.24
6	0.16
7	0.20
8	0.20
9	0.20
10	0.26
11	0.14
12	0.33
13	0.13
14	0.72

15 0.51
16 0.28
17 0.39
18 0.39
19 0.25
20 0.16
21 0.20
22 0.22
23 0.42
24 0.24
25 0.21
26 0.49
27 0.34
28 0.36
29 0.29
30 0.27
31 0.40
32 0.29
33 0.43
34 0.34
35 0.37
Name: A, dtype: float64

Define null and alternative hypotheses

H0 : mean moisture content <=0.35
HA : mean moisture content > 0.35

Decide the significance level Here we select $\alpha = 0.05$.

The sample size for this problem is 36

One sample t test
t statistic: -1.4735046253382782 p value: 0.14955266289815025

Level of significance: 0.05
We have no evidence to reject the null hypothesis since p value > Level of significance
Our one-sample t-test p-value= 0.14955266289815025
In this example, p value is 0.149 and it is greater than 5% level of significance

So the statistical decision is failing to reject the null hypothesis at 5% level of significance.

So at 95% confidence level, there is sufficient evidence to prove that mean moisture present in the Single A is lesser or equal to 0.35 pounds per 100 square feet.¶

1 sample T test for shingles 'B'

0 0.44
1 0.61
2 0.47
3 0.30
4 0.15
5 0.24
6 0.16
7 0.20
8 0.20
9 0.20
10 0.26
11 0.14
12 0.33
13 0.13
14 0.72
15 0.51
16 0.28
17 0.39
18 0.39
19 0.25
20 0.16
21 0.20
22 0.22
23 0.42
24 0.24
25 0.21
26 0.49
27 0.34
28 0.36
29 0.29
30 0.27
31 0.40
32 0.29
33 0.43
34 0.34
35 0.37
Name: A, dtype: float64

Define null and alternative hypotheses

H0 : mean moisture content <=0.35
HA : mean moisture content > 0.35

Decide the significance level
Here we select $\alpha = 0.05$.

The sample size for this problem is 36
One sample t test
t statistic: -3.1003313069986995 p value: 0.004180954800638365

Level of significance: 0.05
We have no evidence to reject the null hypothesis since p value > Level of significance
Our one-sample t-test p-value= 0.14955266289815025
In this example, p value is 0.149 and it is lesser than 5% level of significance

So the statistical decision is failing to able the null hypothesis at 5% level of significance.

So at 95% confidence level, there is sufficient evidence to prove that mean moisture present in the Single B is greater 0.35 pounds per 100 square feet.¶

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

To calculate the p - value and test statistic, we use the scipy.stats.ttest_ind to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.

Ho: population mean for shingles A and = population mean for shingles
H1: population mean for shingles A and != population mean for shingles

tstat 1.2896282719661123
P Value 0.2017496571835306

two-sample t-test p-value= 0.2017496571835306
We do not have enough evidence to reject the null hypothesis in favour of alternative hypothe sis
We conclude that mean population mean for shingles A and B equal.
In this example, p value is 2.2017 and it is grater than 5% level of significance

So the statistical decision is failing to able the null hypothesis at 5% level of significance.

So at 95% confidence level, there is sufficient evidence to prove that mean moisture present in the Single A is equal to Single B